IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Constraining Deep Representations with a Noise Module for Fair Classification

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/1732724 since 2020-03-04T15:17:19Z

Publisher:

ASSOC COMPUTING MACHINERY

Published version:

DOI:10.1145/3341105.3374090

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

04 May 2023

Constraining Deep Representations with a Noise Module for Fair Classification

Mattia Cerrato
mattia.cerrato@unito.it
Università di Torino, Computer Science Department
Torino, Italy

Roberto Esposito
roberto.esposito@unito.it
Università di Torino, Computer Science Department
Torino, Italy

Laura Li Puma
laura.lipuma@intesasanpaolo.com
Intesa Sanpaolo Innovation Center
Torino, Italy

ABSTRACT

The recent surge in interest for Deep Learning (motivated by its exceptional performances on longstanding problems) made Neural Networks a very appealing tool for many actors in our society. One issue in this shift of interest is that Neural Networks are very opaque objects and it is often hard to make sense of their predictions.
In this context, research efforts have focused on building *fair representations* of data which display little to no correlation with regard to a sensitive attribute s. In this paper we build onto a domain adaptation neural model by augmenting it with a "noise conditioning" mechanism which we show is instrumental in obtaining fair (i.e. non-correlated with s) representations. We provide experiments on standard datasets showing the effectiveness of the noise conditioning mechanism in helping the networks to *ignore* the sensible attribute.

ACM Reference Format:
Mattia Cerrato, Roberto Esposito, and Laura Li Puma. 2020. Constraining Deep Representations with a Noise Module for Fair Classification. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC'20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3341105.3374090

1 INTRODUCTION

Artificial intelligence in general, and machine learning in particular, is becoming a pervasive technology adopted by large corporations as well as by small businesses [7]. This growing interest in these technologies is raising increasing concerns about their fairness, motivated by the fact that they could possibly be leveraged to perpetrate and justify discriminating behavior.

In this context deep learning techniques appear to be at odds with the aforementioned trend. Deep learning models have shown impressive results on many pattern matching and machine perception problems [4] and, as a result, businesses and institutions have shown a growing interest in employing neural networks to further automate and innovate their decision-making process; however, a longstanding problem with deep neural architectures is their opacity and sheer number of trainable parameters.

One possible way to cope with this issue, in absence of full explainability, is to ensure *fairness* in the model. Defining a classifier's fairness in a precise, formal way has received considerable attention[5, 8–10]. In this work we leverage neural networks to learn a non-linear mapping from the original data space into a feature space in which information about the sensitive features s is absent. We argue that such a representation is inherently fair.

Instead of explicitly optimizing for a given discrimination measure as in [9], we follow Xie et al. [8] and employ two sub-networks Y and S which are optimized to predict the target and sensitive variables respectively (see Figure 1). Parameters of the model are optimized so that the sensitive attribute is predicted badly, while still allowing for accurate predictions of the target variable.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SAC '20, March 30-April 3, 2020, Brno, Czech Republic
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6866-7/20/03.
https://doi.org/10.1145/3341105.3374090

2 MOTIVATION

Building fair representations through neural networks has been recently proposed in [2, 5]. In these works the neural network is fed with some input data and is trained to learn a representation that discards all information regarding some attribute. Our model is based on Ganin et al.'s work [1, 2] and tries to achieve decorrelated representations via a loss function: $L = L_y - \lambda L_s$ built as the combination of a term L_y that penalizes errors over the target variable y, and a negative term L_s that penalizes accuracy over s, Ganin et al. apply their *Domain Adversarial training* framework to various domain adaptation datasets, showing state of the art performance.

In domain adaptation, learning algorithms are evaluated on their ability to learn features that adapt to other similar, related datasets; on the other hand, state of the art fair classification models produce representations which have little to no correlation w.r.t. the sensitive attribute s. Therefore, we argue that simply employing domain adaptation algorithms to the fair classification context may not result in optimal performances w.r.t. the fairness of the obtained representations. Our experiments in Section 4 indeed show that the standard Domain Adversarial learning algorithm does not guarantee the removal of all information about the sensitive attribute. In the following section we introduce our noise conditioning module which is instrumental in obtaining truly fair representations.

3 PROPOSED MODEL

Let us consider a dataset $D = \{(x_i, s_i, y_i), i \in \{1 \dots N\}\}$, where $x_i \in \mathbb{R}^n$ are vectors describing non sensitive attributes of our problem; $s_i \in \mathbb{R}^m$'s are vectors describing sensitive attributes, and y_i are the target values.

We aim at training a neural network model R so that the output of $R(x)$ can be used as a useful and fair representation for building classifiers predicting y values. We say that a representation r is fair iff s cannot be predicted using only r and we say that it is useful iff y can be accurately predicted.

To achieve this goal, as proposed in Ganin et al. [1, 2], we leverage an auxiliary neural network Y to predict the y variable, and another auxiliary network S to predict the s variable. In the following, θ_Y, θ_S, and θ_R are the parameters for the Y, the S and the R models respectively.

The network R is trained so to maximize the performances of network Y while minimizing the performances of S. Each iteration step optimizes S so to improve in predicting the sensitive variable, optimizes Y so to improve the prediction of the target variable, and optimizes R so to build the best possible representation for Y and the worst possible for S.

The overall training objective can therefore be written as follows:

$$\hat{\theta}_R, \hat{\theta}_Y, \hat{\theta}_S = \arg \min_{\theta_R, \theta_Y} \left[L(y, Y(R(x))) + \lambda \max_{\theta_S} L(s, S(R(x))) \right]$$

where λ is a "fairness importance" parameter employed to balance the importance of decorrelating with respect to s against accurately predicting y. As it should be apparent from the formulation, lower λ values would make the system more tolerant of unfair behaviors (setting it to 0 makes the system behave as a regular neural classifier predicting y).

We note that the optimization problem is complicated by the interaction between the minimization over parameters θ_R with the inner maximization which uses the R network based on those parameters. As suggested in [1, 2], the back propagation algorithm can be easily modified to cope with this objective by changing the sign of the gradients on θ_S after updating the weights of network S (so that the the network R is updated using the *inverted gradient*, thus worsening the performance over S). Figure 1 gives a graphical representation of the model.

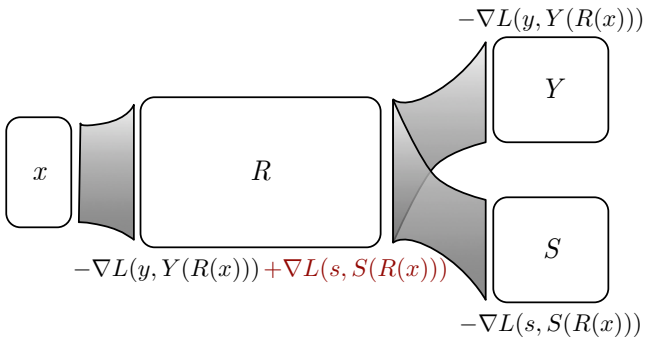

Figure 1: Fairness model. Adjacent to each component of the network we report the gradients used to update the parameters of the component. Note that the direction of $\nabla L(s, S(R(x)))$ is reversed when applied to R.

The model presented so far is equivalent to the one introduced by Ganin et al. [1, 2]. However, as discussed in Section 2, decorrelation with respect to the sensitive variable can prove to be very difficult since it requires to combine many possibly meaningful features to mask and dampen signals correlated with s. To overcome this difficulty, we propose a new noise layer that simplifies the task of decorrelating with respect to s.

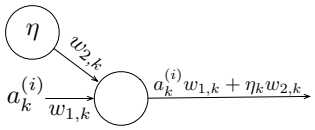

Figure 2: Component k of a noise module inserted after layer i. The number of components k is equal to the number of neurons in layer i. Each neuron's activation, a_k^i, is multiplied by a weight $w_{1,k}$ which can be set to dampen or augment the feature's value; the amount of noise which is added to the same neuron's activation is learned via a separate parameter $w_{2,k}$.

Given a layer i with outputs $(a_k^{(i)})_{k=1}^n$, a noise layer at level $i+1$ computes the outputs $a^{(i+1)}$ as it follows:

$$a^{(i+1)} = a^{(i)} \odot w_1 + \eta \odot w_2$$

where $\odot$ is the Hadamard product, w_1 and w_2 are vectors of learnable weights, and η is a source of noise providing random vectors in $\mathbb{R}^n$. Figure 2 shows the neural module responsible for computing the k-th component of the a^{i+1} output. This parametrization provides the network with an efficient feature-wise mechanism for dampening problematic features which are correlated with s without resorting to searching for combinations with other attributes; at the same time, the addition of an adjustable amount of additive noise can help lowering the amount of mutual information between the learned representations $\hat{x}$ and the sensitive attribute s.

An issue with this approach is that a fundamental assumption of the gradient descent algorithm is that the employed loss function $\mathcal{L}$ is a function of the input data: this assumption is violated if one samples from a distribution at each forward pass of the network, as $\mathcal{L}$ would not be functional (univalent)[1]. We circumvent this problem by only sampling once, at the start of the learning process; while the obtained samples have no correlation with s and can then be employed to partially mask the value of a feature, they are fixed and therefore the functional property of the loss function still holds.

4 EXPERIMENTS

We evaluated our model by running experiments on "fair" classification datasets widely employed in the literature, namely the Adult, German, Bank (from the UCI ML repository [6]) and COMPAS [3] datasets. An archive containing the results of all experiments, as well as the software needed to replicate the experimentation (or to make new experiments) can be downloaded from https://github.com/ml-unito/fair-networks.

The choices for the network architectures follow the encoder network sizes employed by Louizos et al. [5], which were motivated by referring to the sizes of the datasets. Specifically, for the Adult and Bank dataset the networks R, Y, and S are constituted by a single hidden layer with 100 neurons. As for the German and COMPAS datasets, R, Y and S have a single hidden layer with 60 neurons.

In order to pick the best λ value, we repeated the learning process using few different values (namely we experimented with $\lambda \in \{0, 0.5, 1, 2\}$) and picked the most promising one using a simple

[1] That is, $\mathcal{L}$ would not longer be a function, as it would associate different outputs to the same input.

fairness/predictiveness tradeoff evaluation metric, evaluated on a holdout set, that takes into account both the loss in predictivity and the loss in fairness. Lastly, we used the representations to train four general purpose classifiers on the obtained decorrelated representations $\hat{X}$: logistic regression, random forests, support vector machines with Gaussian kernels, and decision trees. The resulting classifiers have been used to predict both the s and y variables over an independent test set. In the following we will focus on the results from logistic regression and random forests since the other results provide similar insights. We set the hyperparameters for the Fair Variational Autoencoder following Louizos et al.'s choices [5].

We observe that our methodology leads to representations which are both fair and discriminative on all datasets, as visible in Figures 3 through 6. In these figures, the solid line refers to the accuracy achievable on the original representation for the data when predicting y, while the dotted line indicates the random chance accuracy for s. Thus, a classifier trained to predict s with a perfectly fair representation would display performance which matches the dotted line, while the performance of a classifier trained to predict y with a representation which has not lost any of its discriminative information would match the solid line.

On the Adult dataset (Figure 3), the fair representations have achieved true invariance w.r.t. s, being close or slightly below the random chance baseline. On the other hand, the standard adversarial strategy is unable to account for all the correlations between the attributes and the sensitive attribute. The Variational Fair Autoencoder achieves representations with a similar level of fairness when compared to our methodology, at the cost of reduced performance when predicting y. Experiments on the COMPAS data provide similar insights. The German dataset displays high levels of class imbalance which lead to non-discriminative representations by all strategies we tested; however, fairness was still preserved. As for the Bank dataset, all of the methodologies employed are able to learn consistently fair and discriminative representations.

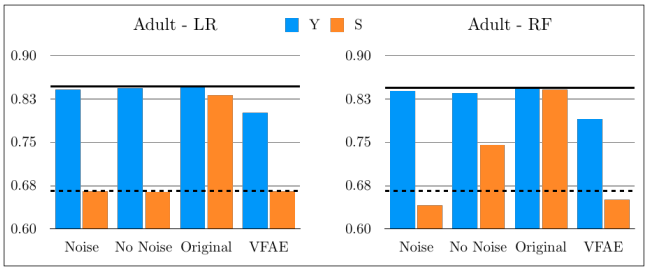

Figure 3: Results on the Adult dataset. In this figure and all the following ones, the solid line refers to the accuracy attained on the original representation when predicting y, whereas the dotted line represents the majority class rate for s. Under our definition, a perfectly fair representation does not allow for performances higher than the dotted line baseline when predicting s.

5 CONCLUSIONS

In this work we have shown how to augment the Domain Adversarial Learning algorithm by Ganin et al. [2] with a noise conditioning layer. Experimental results show that our contribution is instrumental in achieving truly fair representations.

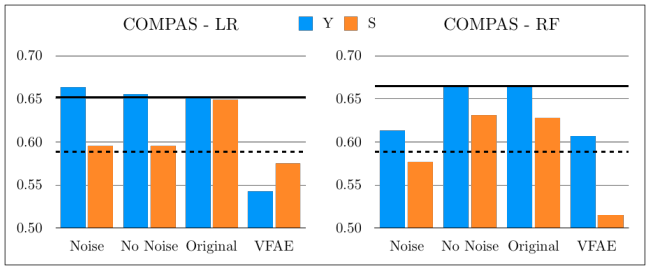

Figure 4: Results on the COMPAS dataset

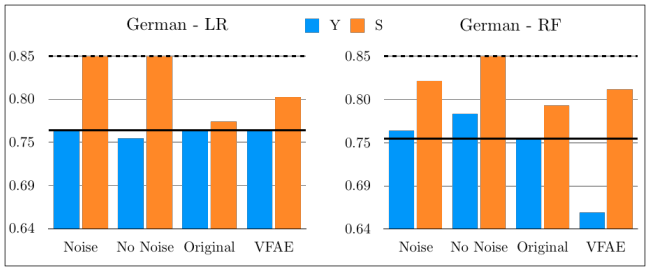

Figure 5: Results on the German dataset

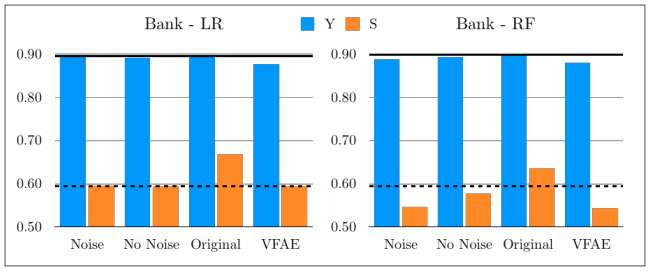

Figure 6: Results on the Bank dataset

REFERENCES

[1] Ganin, Y., Lempitsky, V.: Unsupervised Domain Adaptation by Backpropagation. In: Proceedings of the 32nd International Conference on Machine Learning (2015)
[2] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
[3] Julia Angwin, Jeff Larson, S.M., Kirchner, L.: Machine Bias (2016)
[4] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436 (2015)
[5] Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.S.: The Variational Fair Autoencoder. CoRR **abs/1511.00830** (2015)
[6] Newman, C.B.D., Merz, C.: UCI repository of machine learning databases (1998), http://archive.ics.uci.edu/ml/index.php
[7] Nitin Mittal, David Kuder, S.H.: AI-fueled organizations: Reaching AI's full potential in the enterprise. Deloitte Insights (January 2019)
[8] Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G.: Controllable invariance through adversarial feature learning. In: Advances in Neural Information Processing Systems 30, pp. 585–596 (2017)
[9] Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web (2017)
[10] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning. pp. 325–333 (2013)